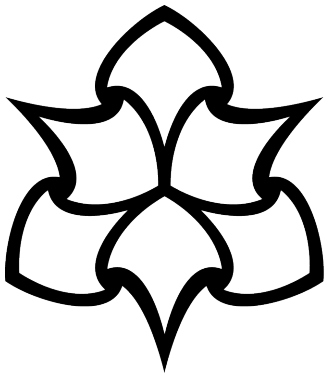

applied sciences

MDPI

Article

Leveraging AI and Machine Learning for National Student Survey: Actionable Insights from Textual Feedback to Enhance Quality of Teaching and Learning in UK's Higher Education

Raheel Nawaz [1,*], Quanbin Sun [2], Matthew Shardlow [3], Georgios Kontonatsios [4], Naif R. Aljohani [5], Anna Visvizi [6,7] and Saeed-Ul Hassan [3,*]

1 Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Manchester M15 6BH, UK
2 Department of Computing & Data Science, Birmingham City University, Birmingham B4 7XG, UK; quanbin.sun@bcu.ac.uk
3 Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, UK; m.shardlow@mmu.ac.uk
4 Department of Computer Science, Edge Hill University, Ormskirk L39 4QP, UK; georgios.kontonatsios@edgehill.ac.uk
5 Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; nraljohani@kau.edu.sa
6 Effat College of Business, Effat University, Jeddah 21551, Saudi Arabia; avisvizi@gmail.com
7 Institute of International Studies (ISM), SGH Warsaw School of Economics, Al. Niepodległości 162, 02-554 Warsaw, Poland
* Correspondence: r.nawaz@mmu.ac.uk (R.N.); s.ul-hassan@mmu.ac.uk (S.-U.H.)

Citation: Nawaz, R.; Sun, Q.; Shardlow, M.; Kontonatsios, G.; Aljohani, N.R.; Visvizi, A.; Hassan, S.-U. Leveraging AI and Machine Learning for National Student Survey: Actionable Insights from Textual Feedback to Enhance Quality of Teaching and Learning in UK's Higher Education. *Appl. Sci.* **2022**, *12*, 514. https://doi.org/10.3390/app12010514

Academic Editors: Jenny Pange and Zoi Nikiforidou

Received: 25 November 2021
Accepted: 29 December 2021
Published: 5 January 2022

Abstract: Students' evaluation of teaching, for instance, through feedback surveys, constitutes an integral mechanism for quality assurance and enhancement of teaching and learning in higher education. These surveys usually comprise both the Likert scale and free-text responses. Since the discrete Likert scale responses are easy to analyze, they feature more prominently in survey analyses. However, the free-text responses often contain richer, detailed, and nuanced information with actionable insights. Mining these insights is more challenging, as it requires a higher degree of processing by human experts, making the process time-consuming and resource intensive. Consequently, the free-text analyses are often restricted in scale, scope, and impact. To address these issues, we propose a novel automated analysis framework for extracting actionable information from free-text responses to open-ended questions in student feedback questionnaires. By leveraging state-of-the-art supervised machine learning techniques and unsupervised clustering methods, we implemented our framework as a case study to analyze a large-scale dataset of 4400 open-ended responses to the National Student Survey (NSS) at a UK university. These analyses then led to the identification, design, implementation, and evaluation of a series of teaching and learning interventions over a two-year period. The highly encouraging results demonstrate our approach's validity and broad (national and international) application potential—covering tertiary education, commercial training, and apprenticeship programs, etc., where textual feedback is collected to enhance the quality of teaching and learning.

Keywords: National Student Survey (NSS); Education for Sustainable Development (EDS); AI for education; higher education policy making; intervention strategies

1. Introduction

Curriculum and testing have a significant impact on the lives and careers of young people. Decisions made by schools influence their students' potential outcomes and chances, and the delayed effects of public examinations and evaluations are perhaps more critical [1]. More recently, it has been argued that evaluation methods in higher education require a

change to include elements within the university didactic assessment strategies, such as co-assessment and self-assessment [2,3].

Nevertheless, it is currently widely accepted that the systematic collection and analysis of student feedback constitutes an important quality assurance and enhancement exercise in teaching and learning in higher education. Nationwide student feedback surveys, such as the UK's National Student Survey (NSS) [4], have become the standard in many countries, and their results used to inform the development of improved teaching interventions and practices [5,6]. The NSS is intended to help students make decisions about where to study and to assist institutions' planning and quality improvement methods, as well as to offer a measure of public accountability. Despite this systematic collection and analysis of student feedback, the evidence [7,8] suggests that, over the past decade, there has been only an insignificant increase in overall student satisfaction. This is attributed to the fact that higher education institutions fail to address in a timely manner the negative aspects of students' learning experiences that are recorded in student satisfaction feedback reports [9,10].

While its analysis provides valuable insights into students' thoughts and perceptions, a number of scholars [11–13] argue that Student Evaluations of Teaching (SET), widely used in academic personnel decisions as a measure of teaching effectiveness, cannot accurately measure the effectiveness of teaching and learning; therefore, it does not constitute a reliable indicator of educational quality. For example, one study [14] shows that the more objective the evaluation process adopted by SET, the less likely it is to relate to the teaching and learning benchmark evaluation standards. Another suggests that there is little or no correlation between SET and teaching effectiveness and student learning [15]. In other words, a wide range of studies have investigated the effectiveness of SET, yet have been unable to indicate specifically how student learning is enhanced.

Furthermore, the vast majority of existing SET studies use Likert-scale responses to capture student feedback, yet, by its very nature, the response to a close-ended question cannot provide as detailed and as in-depth information as that to an open-ended question [16–18] and may sometimes result in an ambiguous finding [19]. For example, one study that aimed to analyze student responses to the NSS 2005/07 open-ended questions [20] revealed that the most frequent issue raised in students' free-text responses pertains to the 'Overall quality of teaching', reporting that the NSS's open-ended questions help to improve the overall response rate. Such findings could not be easily achieved using closed questions due to the necessity of a long list of options.

Despite the many advantages of open-ended questions, the manual analysis of free-text responses requires a considerable expenditure of time and money. As a result, most studies analyze only a small amount of data (in common with other qualitative research). For example, Deeley et al.'s study [21] on exploring students' dissatisfaction with assessment feedback due to recruitment difficulties featured just 44 participants at a college of 5000 undergraduates; MacKay et al. [22] applied social identity theory to their institutional NSS free-text data, processing just 60 responses in total; and, due to time and cost constraints, Richardson, Slater, and Wilson [20] analyzed only 8% of a total of 10,000 free-text responses in their NSS study. Langan et al. [23] processed more than a thousand text comments in their study of the coherence of NSS questions' ratings; although this sample was bigger, the work included only categorization, not in-depth qualitative text analysis. Conversely, this study aims to use text mining to demonstrate a way to analyze a large number of qualitative data (e.g., students' free-text responses) without vastly increased resources (i.e., human effort and time).

Our contributions are as follows: At first, the study proposes a novel automatic analysis framework that can be used to automatically mine the text responses to open-ended questions in student feedback questionnaires. It aims at improving teaching practices and fostering positive learning experiences in higher education through the adoption of machine learning methods. It employs advanced text mining techniques to automatically identify the important problems, issues, and recommendations in students' open feedback on a large scale. The second contribution is the applications of both supervised learning (classification)

and unsupervised learning (topic modeling) to automatically analyze and interpret students' responses to the open-ended questions in the NSS. The third and foremost contribution of this paper is implications of the models on real world data as a case study—comprising of four academic years of NSS (2014–2017) at a northwest UK university. Subsequently, the identified issues were developed into teaching interventions (two at Level 5, and two at Level 6). As a result, the responses to these teaching interventions were collected via the department's end-of-module surveys. The evaluation results showed a significant improvement in student satisfaction (22% higher scores than average), demonstrating the effectiveness of our teaching interventions and the underpinning machine learning methods. Finally, we conclude that the proposed automatic analysis framework was effective in this case and could be applied in the future to substantially reduce the time and cost demanded in practice by this process (e.g., by institution policy and decision-makers). Furthermore, because the information extraction model in the free-text responses those issues and topics of use for a variety of other purposes, our automatic text analysis framework can be easily adopted to wider studies and can be used to track student feedback, influencing institutional policies.

The rest of the paper has been organized as follows: Section 2 presents a detailed review of relevant studies on student evaluation methods and text mining approaches. Section 3 presents detailed methodology, followed by case study in Section 4. Finally, Sections 5 and 6 discuss implications of case study and conclusions, respectively.

2. Literature Review

2.1. Student Evaluation of Teaching

The student evaluation of teaching, or SET, a kind of customer satisfaction survey for higher education, has been widely accepted as a standardized evaluation exercise in North American, UK, and Australian higher education [13]. SET provides a reliable means of assessing its quality of teaching and learning [24] and has been used to: (a) provide benchmark data to compare the quality of teaching and learning systematically across institutions [25]; (b) inform undergraduate applicant choices [26]; and (c) make recommendations to individual institutions on enhancing and improving the quality of their teaching practices [20].

SET is normally conducted in the form of a questionnaire with a list of ordinal-scale questions (for example, 'Rate your overall satisfaction with the course'), using a Likert scale for the response. Students' responses are then analyzed by a wide range of methodologies. SET's reliability as a teaching and learning quality indicator has been debated among pedagogical researchers [27,28], especially regarding whether an ordinal scale can capture quality in teaching and learning, which is considered to be a multidimensional attribute [29–31].

To provide a more comprehensive measurement, many models take multilevel factors into consideration. For example, Toland and De Ayla [32] use a three-factor model to mitigate the correlation effect on students' 'sharing common perceptions to their teachers'; Macfadyen et al. [33] conducted multilevel analysis (combining simple logistic regression and multilevel linear modeling) to test various factors of students' response/non-response to SET; and Rocca et al. [34] demonstrate the effectiveness of adopting an integrated approach (combining IRT and multilevel models) to reveal a student's characteristics (latent traits).

2.2. Text Mining

Text-mining algorithms may take one of two broad approaches: supervised [35] or unsupervised [36]. A high-level overview of a supervised text-mining method is illustrated in Figure 1. In a supervised text-mining setting, a human coder is first required to annotate/label manually a sample of the documents against a pre-defined list of categories. The text-mining algorithms are then trained on the manually annotated samples and learn to associate the textual content of a document with its underlying label (e.g., textual content expresses positive/negative sentiment, or relevance to politics/sports/news/technology).

Once the training process is complete, the algorithm is used to categorize any unlabeled documents automatically to the same pre-defined list of categories. A supervised algorithm can substantially reduce the time needed to analyze a large document collection, whereby a human coder manually categorizes a sample of the documents (normally a small percentage of the whole data), and the remaining unlabeled ones are automatically categorized by the trained text-mining algorithm [37].

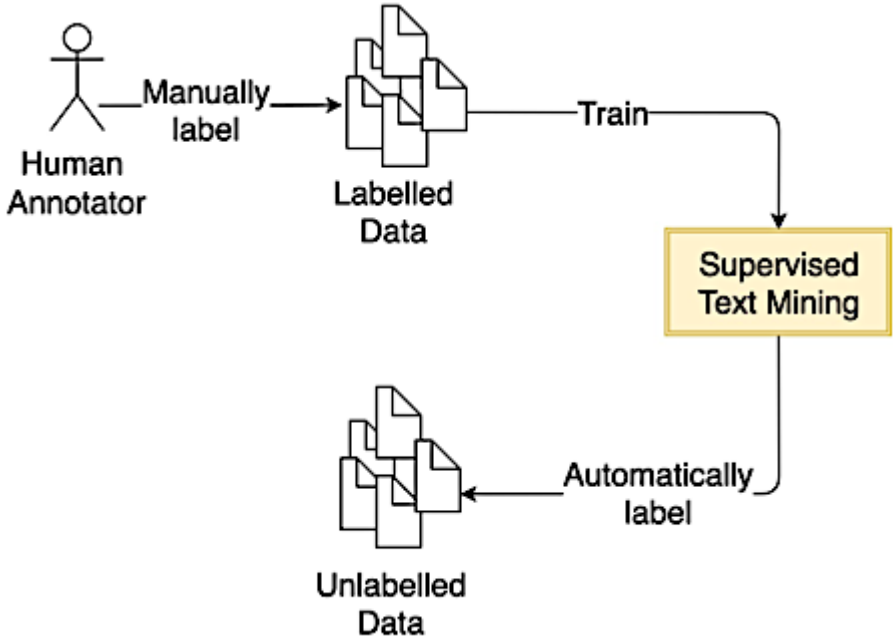

Figure 1. High-level overview of a supervised text-mining approach.

By contrast, unsupervised approaches (illustrated in Figure 2) aim to identify the hidden patterns in a collection of documents (e.g., clusters of semantically similar documents or latent topics) and do not involve a training phase (no need for manual labeling), so can be readily applied to any unlabeled document collection. Unsupervised text-mining methods are widely used to facilitate exploratory analyses of document collections by extracting information that is not immediately evident to experts [38,39]. For example, the most commonly used unsupervised training approaches use document clustering [40] or topic modeling [41].

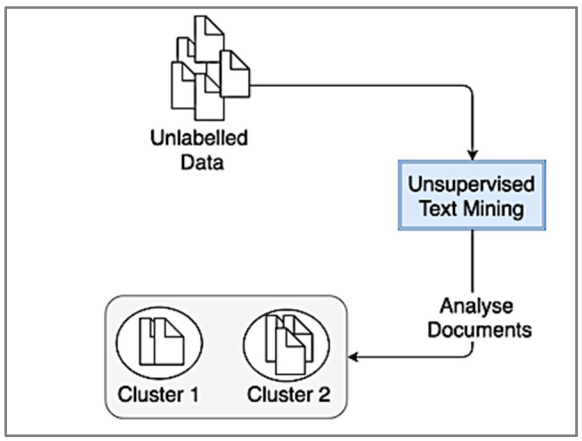

Figure 2. High-level overview of an unsupervised text-mining approach.

Systematic analysis of open-ended survey responses involves a considerable workload in terms of manual annotation; therefore, open-ended responses are often excluded from

survey experiments [42–44]. Nonetheless, survey researchers recognize the importance of analyzing open-ended data, pointing out that such free-text responses to open-ended questions record detailed and useful information that may not be captured by close-ended questions [45,46].

To reduce the workload associated with processing open-ended questions, several studies have explored the use of text mining to automate the underlying analysis process. Reference [47] presented one of the earliest text-mining approaches to the analysis of a dataset of open-ended responses. In their study, the data were collected from a survey that investigated employees' work-related perceptions of working in a large corporation. A dictionary-based text-mining method was used to assign a sentiment score automatically to each free-text response (a high score indicating a positive sentiment, and a low score a negative sentiment). The results achieved a high correlation between the automatically computed sentiment scores of the survey's responses to the open-ended questions and the quantitative scores of responses to its close-ended questions. Reference [43] used an unsupervised text-mining method, namely a term-clustering model, to identify thematically the coherent groups of the terms used in responses to open-ended survey questions. For evaluation purposes, their method was applied to a large-scale dataset of approximately 2000 free-text responses about consumers' preferences. The results showed that the method was able to identify informative and meaningful clusters of the terms that were discussed in the free-text comments.

In another study [44], a statistical topic-modeling method (i.e., an unsupervised text-mining approach) was developed to analyze the topical content of open-ended data automatically. The topical content of open-ended data was represented by a finite set of topics and each, in turn, was represented by a finite set of words. Moreover, to reflect its importance within the open-ended responses, each topic was assigned a weight. This study further demonstrates that the topic-modeling method is able to reveal topics that are semantically similar to hand-coded categories (although some deviations were observed). However, a limitation of such an approach is that the underlying topic-modeling method is totally unsupervised; thus, it naturally ignores any readily available hand-coded categories; therefore, the automatically computed topics may not align perfectly to the humanly annotated (i.e., ground truth) categories.

To address the above-mentioned needs in open-text analysis and the limitations of existing text-mining approaches, we developed a novel method that uses both a supervised and an unsupervised component to analyze students' open responses automatically in order to design effective teaching interventions to enhance the student learning experience. To the authors' knowledge, the study represents the first attempt to investigate the use of text-mining methods to analyze student responses to open-ended questions automatically, as well as to reinforce the teaching activities with those outcomes and evaluate their effectiveness.

3. Methodology: Information Extraction and Teaching Intervention

The study was conducted in two distinct phases, both with several components:

- Phase 1—Information Extraction
  - Annotation Scheme Development: to develop a set of fine-grained categories for students' free-text responses, to be used in analysis.
  - Automated Test Response Categorization: to apply state-of-art supervised machine learning methods in classification to categorize students' free-text responses automatically, using the annotation scheme developed earlier.
  - Issue Subtraction and Summarization: to apply state-of-the-art unsupervised machine learning methods in topic modeling to automatically identify the key issues in each response category and the top-10 issues, taking into account the institutional context.
- Phase 2—Teaching Intervention

- Design: to make an action list of teaching interventions to address the issues identified above.
- Implementation: to implement these actions in selected modules.
- Evaluation: to conduct an end-of-semester survey in each selected module and evaluate its results.

3.1. Annotation Scheme Development

The first stage was to develop an annotation scheme to categorize students' free-text responses, not only to provide a better understanding of the areas that concerned them but to enhance the performance of the later text-mining methods to identify issues automatically. Although there is no existing scheme for text mining (as this study is considered to be the first attempt automatically to analyze NSS open-text responses), Richardson et al.'s [20] findings on NSS free-text data provide a good starting point. After manually analyzing about a thousand responses (both positive and negative) to the NSS open-ended questions, Richardson's study devised 29 categories to cover the whole range of data, with the top-10 most common categories covering 80% of the data. In our study, we too developed an annotation scheme with 10 categories, as shown Table 1.

Table 1. Annotation scheme of problem statement categories used in this study and how it covers the NSS categories.

Category	Proposed Annotation Scheme	Covers NSS Core Questionnaire Categories
1	Overall quality of teaching	The teaching on my course Personal development *
2	Overall level of support	Academic support Learning opportunities **
3	Assessment and feedback	Assessment and feedback
4	Organization, management, and responses	Organization and management Student voice **
5	Learning resources	Learning resources
6	Teaching materials and curricula	N/A
7	Placement and employability	N/A
8	Student life and social support	Learning community **
9	Overall dissatisfaction	N/A
10	Positive or general statements	Overall satisfaction

* Category removed after 2016; ** Category added since 2017.

Our annotation scheme (Table 1) provided full coverage of the NSS core questionnaire categories (The NSS questionnaire was revised in 2017, after it was first introduced in 2005. The National Student Survey website (https://www.thestudentsurvey.com, accessed on 20 August 2021) provides only the latest version. The pre-2016 version can be found available online: http://www.bristol.ac.uk/academic-quality/ug/nss/nssqs05-16.html/ (accessed on 20 August 2021)). This is because, upon manually examining 10% of the free-text responses in our NSS dataset (used as case study from a UK university, consisting of 4400 responses 2014–2017), we found that because the NSS questionnaire uses a Likert scale that provides no additional fields for further comments, rather than to discuss new topics the students often used any opportunity for a free-text response to elaborate on their thoughts. We added three additional categories (Categories 6–8) for more specific topics, based on our examination of our NSS dataset, as well as Richardson et al.'s findings [20]. Although these open-ended questions invited negative comment (positive comments were already covered), we found that a large proportion of responses (29%) were either positive or general, for instance, 'Overall, it's been a good course' or 'No, negatives at this time'. We considered that this was due to students tending to want to start with a positive statement

or perhaps to think that comment was mandatory. Therefore, since we were targeting negative feedback, we created Category 10 for those statements that were nonessential to our study.

We used the annotation scheme that we had developed to categorize about 10% of the responses in our NSS dataset. The distribution of negative statements is shown in Figure 3. Proving its effectiveness and validity, the proposed scheme is well balanced (apart from 'Student life and social support', a category that was included since the impact of social support is often overlooked, and action to improve student wellbeing needs to be more specific) [48–51].

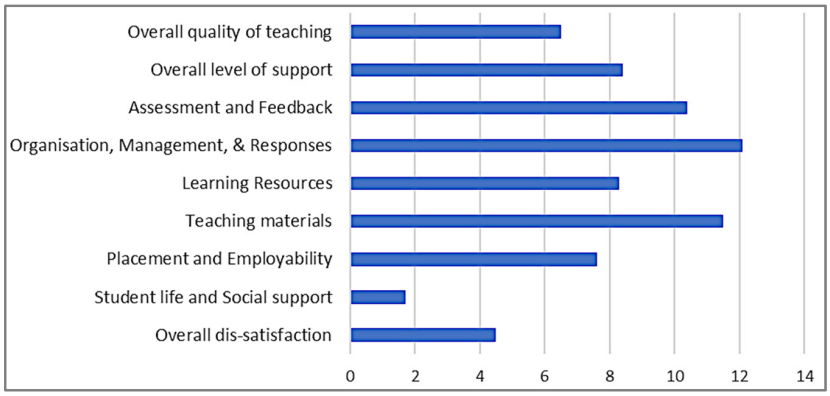

Figure 3. Relative frequency (%) of negative responses per category in the annotation scheme developed.

3.2. Free-Text Analysis—Machine Learning Approach

3.2.1. Framework Overview

This study aimed to identify automatically in the NSS data the issues reported in the free-text responses that make negative comment on the open-ended questions. Figure 4 shows the overall architecture of the proposed text-mining method, consisting of two learning components:

(1) Classification: responses were grouped into smaller categories (e.g., teaching quality, assessment, etc.) to reveal the finer-grained issues to fit the NSS categories, as well as to yield a better performance in topic modeling.
(2) Topic modeling: for each category, important topics were drawn and interpreted into meaningful issues upon which teaching interventions could be developed.

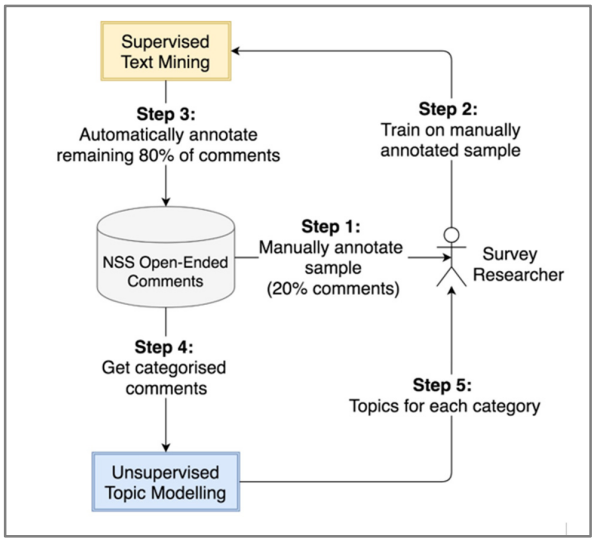

Figure 4. Architecture of the proposed text-mining method to identify automatically the problem statements in the NSS free-text responses.

Specifically, the analysis process was broken down into five steps (Figure 4): Step 1 is initiated by a survey researcher manually annotating a small sample of approximately 20% of the NSS open-ended comments, according to pre-defined categories. Step 2 uses the manually annotated samples to train a classification model to 'learn' to discriminate between the various categories of problem statement. In Step 3, the trained model automatically annotates the remaining 80% of the dataset's open-ended comments. In Step 4, an unsupervised topic modeling method automatically identifies the important topics within each problem statement category. Finally, the automatically generated topics are manually inspected and validated, and the results of analysis inform the development of efficient teaching interventions to address the issues raised by the students.

3.2.2. Automated Test Response Categorization—Classification

To undertake the classification, we divided the full NSS dataset into two subsets:

(1) Labeled data: the text responses that were categorized manually against the annotation scheme, consisting of 20% of the total data, used to train the classifier for the supervised text-mining task.
(2) Unlabeled data: the remaining 80% of the text responses, automatically categorized by the classification algorithm.

In our study, we first evaluated three common classification algorithms (i.e., SVM [52], Multinomial Naïve Bayes (MNB) [53], and Random Forest (RF) [54]) against the labeled data. The best-performing algorithm was then used to categorize the unlabeled data. To evaluate the algorithm performance (i.e., predictive accuracy), the labeled dataset was first divided into a training set (70% of labeled data) and a validation set (30% of labeled data). Secondly, the three classification components were trained on the training set with each algorithm, using the words in the data as predictive features. Thirdly, the trained components were used to categorize the responses in the validation set separately. Lastly, the results were compared to the labeled data in the validation set: the component with the best predictive accuracy was chosen.

3.2.3. Issue Subtraction and Summarization—Topic Modeling

In order to extract latent topics (i.e., to identify the issues in the text responses), for each of the 10 categories we employed the widely used LDA [55], a probabilistic topic-modeling algorithm. Specifically, LDA assumes that each document is a distribution of K latent topics, and each topic is a distribution of M words. In our approach, we used the MALLET toolkit to implement the algorithm, training it for 500 Gibbs sampling iterations and setting the number of latent topics to K = 100. For each topic, we examined the top-seven words (i.e., those that show the strongest correlation to the topic) to summarize the content of that topic. For a better interpretation, the summarization was undertaken jointly by senior lecturers at the institute from which the NSS data came.

As the topic-modeling results were challenging to evaluate [56] and contained only word-based forms, manual interpretation was needed to produce human understandable issues. The results were used later in the teaching intervention.

3.3. Teaching Intervention

To address the issues identified in Phase 1, Phase 2 involved designing, implementing, and evaluating a series of actions to improve students' experience of teaching and learning. For the implementation, we selected four modules run by the Department of Computer Science (at the institute from which the NSS data came): two at Level 6, as their students would subsequently participate in the NSS survey; and two at Level 5 to provide a more comprehensive evaluation. An end-of-module survey was conducted, using as its baseline the previous year's module evaluation results to measure the effectiveness of our teaching intervention. As Phase 2 uses the result of Phase 1, it is introduced and discussed in the later section, 'Phase 2 Design and Result'.

4. Case Study: NSS Data from a UK University

The dataset was collected over four academic years of NSS (2014–2017) at a northwest UK university. In this study, we used only the responses to the question about negative learning experiences, which had 4400 responses.

4.1. Phase 1: Textual Analysis with Machine Learning

4.1.1. Classification of Open-Ended Responses

It should be noted that, in this study, the statement categorization was the performance with an individual sentence, rather than with the whole response. This considers that a single response may consist of statements in multiple categories. For example, Figure 5 shows a response that falls into four problem statement categories. Accordingly, these responses were broken down to 10,328 sentences.

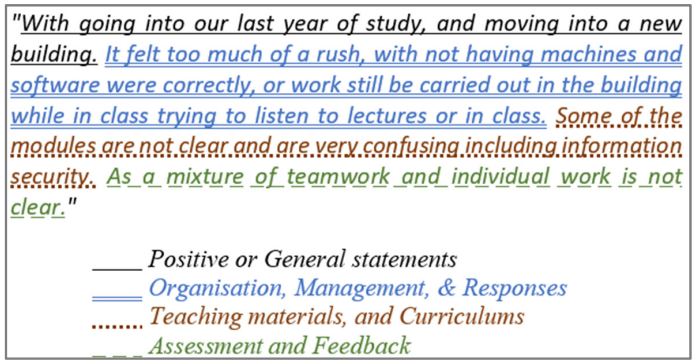

Figure 5. Example of a single student response consisting of multiple problem statement categories.

About 20% of sentences were randomly selected and manually annotated against the annotation scheme (Table 1). As 10% had been done at Stage 1, this annotated another 10% and produced 2000 labeled data. The annotation in both stages was undertaken by two senior lecturers at the institute from which the NSS data came. The unlabeled data consisted of the remaining 8328 sentences, which were automatically analyzed later by text mining.

Furthermore, we trained (with 70% of the labeled data, i.e., 1400 sentences) the text-mining component with the following algorithms: Random Forest (RF); Multinomial Naïve Bayes (MNB); and Support Vector Machines (SVM). Their prediction accuracies (against a held-out 30% of the labeled data) are shown in Figure 6.

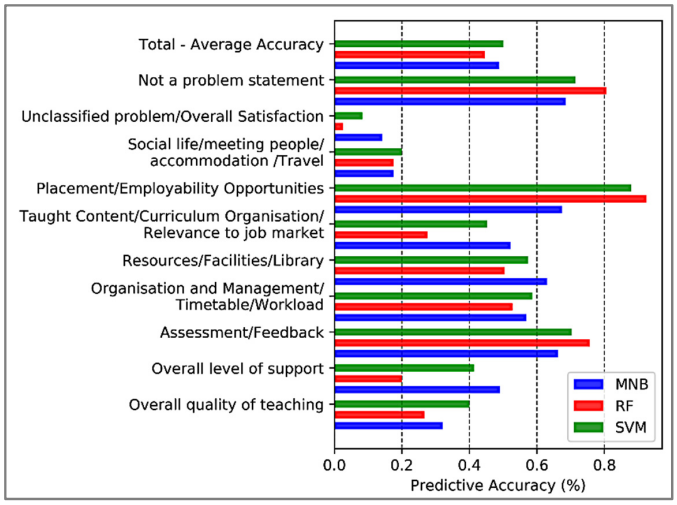

Figure 6. Predictive accuracy (%) of the Multinomial Naïve Bayes (MNB), Random Forest (RF), and Support Vector Machines (SVM) algorithms when applied to the evaluation sample of the NSS dataset.

Overall, the SVM algorithm yielded the best average accuracy (50% macro-average across all classes), although its performance improvement over MNB (+1.3%) was low. It should be noted that the accuracy obtained by the three algorithms varied across the 10 problem statement categories. For example, the 'Placement and employability' category showed the best predictive accuracy of all the categories with the RF algorithm, achieving a robust performance of 92.5%. The high performance in this category could be explained by the fact that many student responses relevant to 'Placement and Employability' contained the word 'placement', which was used by the underlying algorithm as a discriminating feature to categorize such problem statements accurately. Two other categories, namely 'Overall dissatisfaction' and 'Student life and social support', achieved a very low accuracy of approximately 2–14% and 17–20%, respectively, attributed to the limited number of training examples available in these two categories.

As a result, the trained SVM component was chosen to categorize the responses in the unlabeled data. Figure 7 shows the distribution of the statements in each category. Although the classification accuracy of the trained SVM was not high for all classes, we still consider this useful for increasing the amount of silver-standard labeled data that we are able to provide to our topic-modeling algorithm. The results shown in the next section indicate that the topic modeling algorithm was able to produce coherent outputs when using our automatically labeled data.

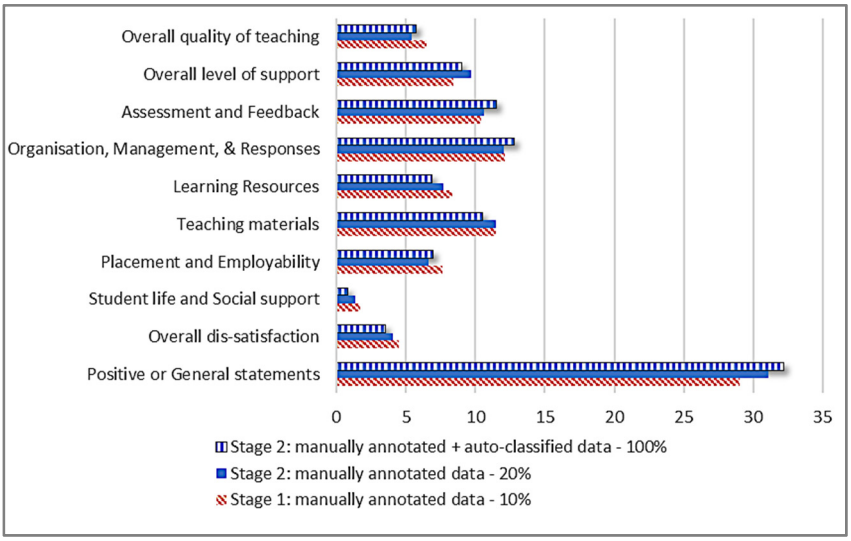

Figure 7. Relative frequency (%) of statements (manually annotated and auto-classified) per category using the annotation scheme developed.

4.1.2. Results of Topic Modeling on NSS Data

This step automatically lists the important topics among the categorized statements. For each category, we applied the LDA topic-modeling method to summarize the top-10 topics. This generated a hundred topics. For brevity, Figure 8 shows the results for the 'Assessment and Feedback' category as an example; the rest are in the Appendix. The first observations were that the topic-modeling method was able to identify thematically coherent topics (with a topic weighting to reflect their relevance) and also represent the underlying category. However, human interpretation was required since each topic was described by a set of seven words, and some topics could be better summarized as a single issue. For example, both Topics 3 and 10 indicated 'Timely feedback to assessment'; Topics 2, 3, and 8 suggested 'Clearer assessment criteria'; and Topics 5, 7, 8, and 9 related to 'Assessment deadlines'.

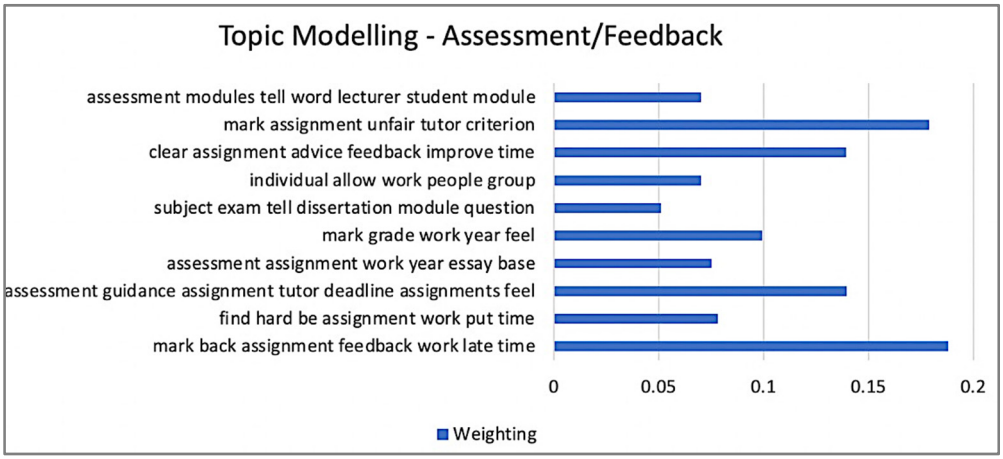

Figure 8. The 10 most important topics in the 'Assessment and feedback' category.

Using human interpretation and a holistic view of a hundred auto-summarized topics, we were able to summarize the top-10 issues (from more than 10 topics, as demonstrated above) across all categories. Although this process was not made fully automatic, we consider that human involvement was far less than that required to process all 10,328 statements manually. Finally, top-10 issues identified are as labeled (I1–I10), which are further discussed to inform the teaching interventions in the next stage.

4.2. Phase 2: Teaching Intervention and Evaluation

4.2.1. Discussion on Issues Identified

The automatically identified topics/issues were considered to reflect the authors' teaching experience in the department and were in line with previous studies [4,10,57]. Specifically, I1, I2, and I3 revealed students' concerns over the teaching and learning materials. We considered that such concerns were caused by insufficient communication rather than an inadequate quality of material, because the department had undergone a rigorous process (including staff training, use of templates and moderation) to ensure the quality of all teaching and learning materials; in addition, it had always received positive feedback from external examiners. Moreover, the students were of diverse backgrounds (i.e., more BTech and converted students than at a traditional university) and whose familiarity with the higher education learning environment varied significantly. For example, the academic writing in the materials may have constituted a barrier to their learning, as revealed by I3, the topic about marking criteria. To address such issues, we suggested that the teaching interventions should focus on providing students with guidance on accessing relevant information and that during the semester each module should have more checkpoints for students (regarding self-assessment and identifying misunderstandings).

I4, I5, and I6 relate to the modules' content. While I4 was a generic description of how the content needs to be more interesting and engaging, I5 and I6 revealed more detail, clearly suggesting that the students would like to gain more practical skills upon successful completion of the various modules. Some might argue that Computer Science is a field that combines both theoretical knowledge and practical skills and that students may not have the ability to judge just how practical a Computer Science module should be; however, it has been reported elsewhere that Computer Science graduates lack practical skills [58] and that work experience is important to them [59]. To address these issues, more practical tasks and real-life examples should be integrated into modules. In addition, when delivering lectures, whenever possible, tutors should explain to the students how the theoretical concepts are applied in industry (ideally, with real-life examples) and how the acquired practical skills contribute to their career.

I7 and I8 refer to the timing and quality of feedback on students' coursework. Both issues initially appear surprising, as the delivery of feedback at the institution follows a

rigorous process in the Computer Science department, especially as, in previous years, it introduced an additional checkpoint (a marking mentor to oversee the process and monitor the quality of feedback). Based on the authors' extensive experience of marking and moderation (over a range of over twenty modules) and the comments by other senior teaching members in the department, we concluded that the feedback given was actually both detailed and constructive, and that some even exceeded the department's requirements; therefore, it was worth exploring possible alternative explanations of these issues.

After careful consideration and discussion with student representatives, it was suggested that both issues may relate to the response time of official feedback (i.e., four weeks). First, for major coursework assignments that are usually submitted at the end of a module, students receive feedback only when the module is completed, after the end of term. Second, for minor coursework assignments that are submitted frequently (e.g., the weekly portfolio), students do not receive feedback before starting their next assignment. The existing mechanism for providing feedback to students is not ideal, as students receive no direct guidance on how to improve their marks, potentially frustrating them and resulting in negative survey responses.

While the official feedback response time cannot be shortened and the existing assessment arrangements cannot be changed, we recommend that, for major assignments, several checkpoints should be introduced to monitor students' progress and provide informal feedback and/or formative exercises as sub-assignment tasks; for minor assignments, informal/oral feedback should be provided and prior to the next assignment there should be discussion in class on common mistakes (this requires tutors to go quickly through all submissions). For all assignments, the marking criteria should be clearly explained and linked to the task, which will help the students to understand better the feedback that is provided.

I9 and I10 relate to the experience of the support that students receive from tutors out of class. In this department, the official response to a student email must be within three working days, and there are designated office contact hours for students. In the follow-up investigation, these policies were clear to all staff members, and there was no evidence of staff violating any policies. While this initial finding contradicts the issues identified, further exploration suggested that the issues could be due to inconsistencies in the level of support provided by tutors. For example, some staff tended to reply to student emails within minutes, and sometimes even outside of office hours, while others tended to respond only during office hours. When a student popped in during non-contact hours, some staff tended to put aside any task that they were working on, while others followed the policy and booked a later appointment with the student. Such varied behaviors gave a false impression to students that every staff member should at all times respond or offer help (i.e., the students were unclear about the policies and took as the norm the instant response), a finding that was later confirmed by the student representatives. To address those issues, it is suggested that, at the beginning of a module, all tutors should explicitly clarify: (a) what to expect when emailing a tutor; (b) how a tutor's office hours work; and (c) the best way of contacting them.

4.2.2. Teaching Intervention: Design and Implementation

Our analysis revealed that some issues in student feedback required additional action (e.g., enrichment of module content by including more practical tasks, more frequent checkpoints, feedback, etc.), while others required only amendments or a new delivery approach (e.g., a better communication or clearer explanation). As a result, we propose the actions shown in Table 2.

The actions (in Table 2) aim to address all the issues identified by the text-mining method. While Table 3 shows detail description of top ten issues, Table 4 provides a detailed overview of how each identified issue.

Table 2. Actions to improve student experience.

	Action	When
A1	An introduction session to the module schedule and timetable	Week 1
A2	Display timetable and key date on a separate page on Blackboard	All weeks
A3	An introductory session to explain how the module is linked to the other modules in the pathway and how it contributes to students' careers	Week 1
A4	A walk through to the Blackboard area and module handbook	Week 1
A5	Detailed explanation of assignment, how it will be assessed, and how the feedback will be provided	Weeks 1–2, Week 11
A6	Embed real-life examples into lectures	Lecture
A7	Link assignments to real-life scenarios	Coursework
A8	Embed gamification into lectures	Lecture
A9	Display the tutor's contact method and time on Blackboard homepage and cover page of each lecture note	All weeks
A10	Try to identify and resolve students' queries and issues within the tutorial session	All weeks
A11	Emphasis how email communication works and promote face-to-face appointments with students	All weeks
A12	Check students' progress on coursework and provide oral feedback and suggestions to improve	Weeks 9–12
A13	Provide a provisional mark to students' coursework drafts by referring to the marking criteria	Weeks 9–12

Table 3. Issue code (I1–I10) and description.

Issue Code	Description
I1	Timetable and schedule are not clear
I2	Assessment date and submission deadline confusion
I3	Assessment criteria to be made clearer
I4	Content to be made more interesting and stimulating
I5	Lack of practical tasks
I6	How to apply skills learnt into practice
I7	Timely feedback to be provided
I8	Feedback not helpful
I9	Ability to contact tutor
I10	Staff email response

Furthermore, four modules were selected for implementation. The intervention was conducted during the subsequent academic year:

- Module A, Level 5: 54 students.
- Module B, Level 5: 4 students.
- Module C, Level 6: 49 students.
- Module D, Level 6: 35 students.

The selection criteria were that the modules at both levels should cover all the student pathways and have more than 30 students. It should be noted that Module B, when it was too late to arrange another, enrolled only four students (because the module choice event was held after the study had been designed). The three other modules were able to cover all pathways.

Table 4. Issues addressed by actions.

Action \ Issue	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
A1	x									
A2	x	x								
A3				x	x	x				
A4						x				
A5		x	x							
A6		x	x				x			x
A7				x		x				
A8					x	x				
A9				x	x					
A10									x	
A11							x	x	x	x
A12							x			x
A13		x					x	x	x	

4.2.3. Evaluation: Discussion on Questionnaire Results

We chose a questionnaire as our evaluation method as it is used by NSS and is an effective way to gather large numbers of students' feedback [60]. We extended the department's end-of-year module survey for the purpose of our evaluation. The survey design is in alignment with the NSS survey, offering five Likert-scale options: Strongly disagree; Disagree; Neither agree nor disagree; Agree; and Strongly agree. The full questionnaire, corresponding issues and actions are shown in Table 5. Questions 1 to 7 are part of the set of standard questions from the department's end-of-year survey. Questions 8 to 18 are the additional ones put to the students in the modules under study. Finally, the surveys were conducted at the end of the spring semester, and the results are shown in Table 6. The satisfaction rate was calculated by combining the 'Agree' and 'Strongly agree' responses (note: the Level 5 result does not include Module B, since only one of its four students returned the questionnaire).

Table 5. Questions used in the survey.

	Question	Related Issue	Related Action
1.	The assessment brief helped me understand what was required from the assessment	2, 3	1, 4
2.	The criteria used in marking assessments have been made clear in advance (e.g., in module handbooks, via Blackboard, in class)	2, 3	4, 5
3.	Feedback supported me in terms of aiding my learning and development	7, 8	10, 12, 13
4.	Tutors/lecturers were supportive	4, 8, 9	8, 9, 10, 12, 13
5.	Formative feedback supported my learning	7, 8	10, 12, 13
6.	Staff were available and easy to contact during office hours and the working week	9, 10	9, 11
7.	Overall, I am satisfied with the quality of the module	All	All
8.	Real-life examples helped me better understand the content of the module	4, 5, 6	3, 6, 7
9.	The number of real-life examples provided in class is adequate	4, 5, 6	6
10.	Real-life examples increased my motivation and engagement with the module	4, 5, 6	3, 6, 7
11.	Marking criteria are stated in a clearer way in the module handbook compared to previous years	3, 7	5
12.	Marking criteria are better explained by the tutor compared to previous years	3, 7	4, 5
13.	In-class feedback on my coursework draft was helpful	6, 7, 8	10
14.	Coursework deadlines were clearly communicated	1, 2	1, 2
15.	Regular check of coursework progress by tutor helped me to better manage my time	7, 8	10, 12
16.	Regular check of coursework progress helped me to improve the quality of my coursework assignment	7, 8	10, 12, 13
17.	Tutor communicated clearly the expected outcomes of the coursework assignments	2, 3, 8	4, 5, 13
18.	Tutor communicated clearly what should be done to improve the quality of my coursework	8	10, 12

Table 6. End of module survey results.

Question Summary		Satisfaction Rate (%)			
		Level 5	Level 6	NSS *	Level 6 against NSS
1.	Assessment brief	85	100	n/a	
2.	Marking criteria	96	94	71	+23
3.	Feedback	63	91	69	+22
4.	Staff support	78	100	73	+27
5.	Formative feedback	68	94	n/a	
6.	Staff contact	74	94	83	+11
7.	Overall	59	100	73	+27
8.	Real-life examples	83	100	n/a	
9.	Real-life examples	66	94	n/a	
10.	Real-life examples	55	83	n/a	
11.	Marking criteria	62	67	n/a	
12.	Marking criteria	58	78	n/a	
13.	In-class feedback	62	94	n/a	
14.	Coursework deadlines	100	94	n/a	
15.	Progress check	68	83	n/a	
16.	Progress check	75	89	n/a	
17.	Staff communication	90	100	n/a	
18.	Staff communication	86	94	n/a	
	Average	74	92	74	+22

* The ratings were calculated by averaging all programmes (weighted by student numbers). Data refers to the previous year and the same department in which the teaching intervention took place. NSS survey results available online: http://www.hefce.ac.uk/lt/nss/results/ (accessed on 20 August 2021).

The results of the Level 6 modules show an average rating of 92%; 13 of 18 questions received 90%+ ratings, and five questions achieved 100%. The results of the Level 5 modules show an average rating of 74%, and three questions achieved 90%+.

As they were part of the NSS questionnaire, for Questions 2, 3, 4, 6, and 7, we listed the department's previous year's NSS scores (an average rating of 74%) to provide a direct comparison: the results for the Level 6 modules show substantial improvement (all responses were rated higher), while Level 5 yielded mixed results. Since the actions were designed for Level 6 students, the survey results indicate a significant improvement in students' experience of teaching and learning in the department: on average, 22% more satisfaction. The Level 5 results are for comparison and will be discussed further in a later section.

5. Discussions and Implications for Decision Making

5.1. Effectiveness of the Teaching Intervention

Overall, the interventions designed are considered to be effective, achieving an average satisfaction rate of 92% and, for Q7 (overall satisfaction), 100%. Note that only Level 6 results are included in this discussion. To measure the effectiveness of each action, we took the average rating of related questions, as defined in Table 6 (the rating for Q7 is excluded, as it is an overall rating). For example, A1 has two related questions—Q1 and Q14—whose ratings are 100% and 94%. Therefore, the rating for A1 is (100 + 94)/2 = 97 (%). The detailed measurements of each action are listed in Table 7.

Table 7. Actions and average rating of related questions.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
Average rating of related questions	97	94	92	93	85	92	92
	A8	**A9**	**A10**	**A11**	**A12**	**A13**	
Average rating of related questions	93	97	92	83	92	100	

It was concluded that most actions were well received by students: 11 actions achieved a 90%+ rating. A5 ('Explanation of how assignment works') showed an 85% rating and A11 ('Explanation of how email contact works') 83%, which appears to suggest that the clarification of existing documentation is less helpful to students. However, A13 ('Provisional marking criteria'), which had a 100% rating, is also about explaining documentation (i.e., the marking criteria). Therefore, we conclude that negative responses were probably caused by how the teaching was delivered rather than by the clarity of documentation and policies. In other words, clarification should be carried out by linking to the context rather than by simply referring to the documents. For example, instead of directly explaining each of the marking criteria in Week 1, it is more effective to link a difficult/key concept to the marking criteria, alongside teaching, and explain how it will be assessed, as this will help students to plan their after-class study better, enhancing their learning experience. To measure the effectiveness of how each issue was addressed, we calculated the average rating of all questions related to that issue, as defined in Table 6 (Q7's rating is excluded). For example, I2 is related to Q1, Q2, Q14, and Q17, whose ratings are 100%, 94%, 94%, and 100%. Therefore, the rating for I2 is (100 + 94 + 94 + 100)/4 = 94 (%). Table 8 shows the measurements of all issues.

Table 8. Average rating of issues and related questions.

Issue	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**
Average rating of related questions	94	97	88	94	92	93	85	92	97	94

While most issues were addressed well and eight of the 10 scored a 90%+ average rating, we would like to discuss further I3 and I7, which were rated at under 90%.

For I3 ('Assessment criteria to be made clearer'), which shows a rating of 88%, the related questions can be categorized into two groups: (a) whether the students feel it is helpful (Q1, Q2, and Q17); and (b) whether the students feel that it is better than in previous years (Q11 and Q12). The former shows an average rating of 98%, and the latter 73%. This suggests that the students are less sensitive to the improvements made to documents than to how the teaching materials were delivered; that is, they prefer explanations about their actual assignments to documents giving the marking criteria. This analysis is in line with the previous finding, indicating that students prefer in-context explanation. As a result, we recommend that future actions should continue to focus on practical and contextual items.

I7 ('Timely feedback to be provided') was considered to be difficult to address. This is because the institution guidelines stipulate that marks and feedback on coursework assignments should be returned to students within four weeks so that, by the time that written formative feedback arrives back to students, it offers them little support for improving their work for their current module. While the feedback's high quality was confirmed by the survey results (Q3, Q5, Q13, Q15, and Q16 had an average rating of 90%), only 73% considered that the feedback was more helpful than in previous years (Q11 and Q12). In particular, 94% (Q13) students praised the in-class feedback, suggesting that further actions should focus on oral and informal formative feedback. This is supported by the 100% rating for Action 13 ('In-class check to a provisional score').

5.2. Comparison of Level 5 and Level 6 Results

The average rating of Level 5 (74%) is much lower than that of Level 6 (92%). Although it can be argued that the teaching interventions were based on Level 6 students' feedback (NSS involves only Level 6 students), this is a highly interesting result that will be discussed from several aspects.

The first question is whether the actions were appropriately implemented at both levels. For finer measurement, we split the questions into three categories: (a) four questions directly asked whether a certain action was useful: Q1 (assessment brief); Q2 (marking criteria explanation); Q8 (using real-life examples); and Q14 (coursework deadline). Level 5 results had an average rating of 91%; (b) four questions related to tutors' supportiveness: Q4 (overall); Q6 (available time); Q17 (expected outcome); and Q18 (improvement). Level 5 results show an average rating of 82%. (c) four questions asked about whether the actions were sufficient: Q3 (feedback); Q9 (example numbers); Q13 (in-class support); and Q16 (in-class check). The Level 5 results show an average rating of 67%.

The above statistics might indicate that, while the actions were helpful, they should be implemented more frequently or that tutors should improve their overall quality of teaching. However, it should be noted that on the chosen Level 5 module the lecturing hours were reduced during the year from four to three per week yet the module content was unchanged (the department made the decision just before the start of the semester, leaving no time to amend the teaching material). We observed that the reduced teaching time resulted in increasing pressure on students and a weaker understanding (the module average mark dropped from the previous year). In view of this negative issue (it is notable that only 53% students considered that they were satisfied with the module), it is considered that the actions were implemented well, but that their effects were overshadowed. To resolve this issue, action should be taken either to increase staffing or to modify the module.

The second question to be asked is whether the actions designed efficiently address the issues raised by the students. As mentioned above, the Level 5 ratings on whether the actions are useful were similar to those on the Level 6 modules, yet the overall rating was much lower. Furthermore, Q10 asked whether the real-life examples increased the motivation', to which students at Level 5 gave 55% while those at Level 6 gave 83%. It seems that such an action ('Embedding real-life examples) had no effect on students' learning experience at Level 5. Bear in mind that the actions were designed by analyzing feedback from final-year students (Level 6), who might have had a different perspective from that of the Level 5 students. We consider that the sample data might not be sufficient to draw a conclusion (i.e., whether an action can improve overall satisfaction), but it was clear that to some extent those actions could improve students' experience (i.e., 91% considered the actions were helpful).

6. Conclusions

6.1. Significance of Using Automatic Textual Analysis

We first measured the efficiency of the proposed automatic framework in terms of how it accelerates the process of analyzing students' responses to the open-ended questions. Specifically, at the stage of assigning the responses (10,328 sentences) into the 10 categories, our classification model required 2000 sentences to be manually processed, saving 80% of the human effort (roughly two weeks' work); at the subsequent stage of further summarizing the key issues within each category, we used topic modeling to weight the top-10 topics for each category, involving only a trivial amount of human effort compared to reviewing and prioritizing the whole dataset manually. Assuming a similar time for human review, using our automatic analysis saved four weeks' work. Furthermore, future studies can apply the trained classification model directly, obviating the need for initial manual categorization.

Second, the efficiency of the proposed automatic analysis can be measured by the quality of the identified issues. As we did not have the resources (two weeks for each stage, as estimated above) to go through the entire dataset, we used an indirect measurement to

validate the identified issues against the existing literature and the authors' local teaching experience (as the responses were from students at the same institute), as discussed in the section 'Discussion on issues identified'. The outcomes were meaningful and were developed into teaching interventions. The survey results show that the students' learning experience improved as a result; it should be noted that, although the average accuracy of classification was relatively low, topic modeling was able to tolerate and produce validated outputs, suggesting that the two-stage analysis framework that was designed is resilient. Topic modeling is designed to identify salient features in a text and is robust to noise introduced by misclassifications.

Third, automatic analysis becomes more efficient on a large scale. This is because the requirement for training data does not increase linearly. In other words, human analysis requires 10 times the hours to handle 10 times the amount of data; however, in text mining, while more labeled samples lead to better performance, the effect of increasing the extent of the training data becomes less apparent when the samples are large enough. In this study, we manually labeled 20% of data and the net number is 2000 sentences. While it is hard to estimate how many samples are required to achieve an optimal performance in auto-analyzing students' open response, it is not difficult to predict that, to analyze a million student responses, there is no need to manually label 20% of the data (0.2 million). This is because the trained model has already achieved a reasonable outcome with 2000 sentences.

6.2. Limitation and Future Work

Due to the limited time and resources, our study has the following limitations.

First, although the text-mining results were automatically generated, human interpretation of the obtained results was still required. In addition, the supervised component of our method showed a relatively low average predictive accuracy of 50% across the 10 pre-defined categories. This was due its inability to produce a large amount of quality training data. Since the aim of this study was to concept-prove the effectiveness of our novel text-mining method to identify the issues expressed in students' feedback and thereafter to design and implement interventions, a small training set was considered sufficient. A future work plan is to collect a larger set of training data for better performance, as well as to improve the quality of the labeled data by adapting a more rigorous annotation methodology, as suggested by other researchers [61,62].

Second, while our intervention and evaluation were implemented in four selected modules at two levels, we understand that the NSS data used in this study provide programme-based feedback. Although the modules were carefully chosen to include students from all programs, it became evident that the evaluation results varied across the modules and levels. For example, the action of 'Embedding real-life examples' was generally well received by Level 6 students but received lower ratings from Level 5 students. It is suggested that, when designing interventions in future, module-level feedback should be taken into account.

Lastly, the survey results showed the positive effect of our intervention, which is considered sufficient to justify the effectiveness of our novel approach. However, statistically speaking, such a result might not be considered to be robust enough to inform the development of university policymaking or teaching guidelines. At the 95% confidence level, the average rating is 92%, suggesting that the results are encouraging but not statistically significant. In the next stage of our study, to obtain a more reliable result, we will encourage more students to participate in the module evaluation survey.

6.3. Summary

In this article, we present a full-length pedagogy study on improving students' experience of teaching and learning, covering all the stages from the initial text analysis, issues of identification, teaching intervention design, and implementation to the final evaluation. The significant improvements in student satisfaction demonstrate the effectiveness of our teaching intervention, further proving the accuracy of the text analysis of students' free-text

feedback. In contrast to traditional manual text analysis, we took a novel machine learning approach, enabling us to analyze the free-text comments on a wider scale (i.e., 4400 NSS responses), taking much less time and human resources. Technically speaking, this automatic analysis framework demonstrates its efficiency in 'closing the gap' between the analysis of student feedback and the implementation of timely teaching interventions. Moreover, we consider that this framework has huge potential to be extended: the information extracted from the free-text responses could be used for a wider range of studies; and the annotation scheme developed in Phase 1 could serve as the baseline for future classification studies. A wider implication of this work is beyond higher education, as the methods employed in this research can be leveraged for case studies from tertiary education, commercial training, and apprenticeship programs, etc., where textual feedback is collected from to improve the quality of teaching and learning.

Author Contributions: Conceptualization, R.N., Q.S., G.K. and S.-U.H.; Data curation, M.S.; Formal analysis, R.N., Q.S., G.K. and S.-U.H.; Investigation, S.-U.H.; Methodology, R.N., Q.S., M.S. and G.K.; Resources, N.R.A. and A.V.; Software, N.R.A.; Supervision, R.N. and S.-U.H.; Visualization, A.V.; Writing—original draft, R.N. and S.-U.H.; Writing—review & editing, M.S., N.R.A. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ortega, J.L.G.; Fuentes, A.R. The accuracy of student's evaluations. Study of their evaluative competences. *REDU Rev. Docencia Univ.* **2017**, *15*, 349–366. [CrossRef]
2. Fuentes, A.R.; Ortega, J.G. An Analytical Perspective of the Reliability of Multi-Evaluation in Higher Education. *Mathematics* **2021**, *9*, 1223. [CrossRef]
3. Botaccio, L.A.; Ortega, J.L.G.; Rincón, A.N.; Fuentes, A.R. Evaluation for Teachers and Students in Higher Education. *Sustainability* **2020**, *12*, 4078. [CrossRef]
4. Higher Education Funding Council for England. National Student Survey Results 2016. Available online: http://www.hefce.ac.uk/lt/nss/results/2016/ (accessed on 20 August 2021).
5. Harrison, R.; Meyer, L.; Rawstorne, P.; Razee, H.; Chitkara, U.; Mears, S.; Balasooriya, C. Evaluating and enhancing quality in higher education teaching practice: A meta review. *Stud. High. Educ.* **2022**, *47*, 80–96. [CrossRef]
6. Paek, S.; Kim, N. Analysis of Worldwide Research Trends on the Impact of Artificial Intelligence in Education. *Sustainability* **2021**, *13*, 7941. [CrossRef]
7. Shah, M.; Nair, C.S. The changing nature of teaching and unit evaluations in Australian universities. *Qual. Assur. Educ.* **2012**, *20*, 274–288. [CrossRef]
8. NSS. National Student Survey. 2015. Available online: http://www.hefce.ac.uk/lt/nss/results/2015/ (accessed on 20 August 2021).
9. Nair, C.S.; Pawley, D.; Mertova, P. Quality in action: Closing the loop. *Qual. Assur. Educ.* **2010**, *18*, 144–155. [CrossRef]
10. Symons, R. *Listening to the Student Voice at the University of Sydney: Closing the Loop in the Quality Enhancement and Improvement Cycle*; Australian Association for Institutional Research Forum: Coffs Harbour, Australia, 2006; Available online: http://www.aair.org.au/app/webroot/media/pdf/AAIR%20Fora/Forum2006/Symons.pdf (accessed on 14 September 2021).
11. Boring, A.; Ottoboni, K. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *Sci. Res.* **2016**, 1–11. [CrossRef]
12. Alauddin, M.; Kifle, T. Does the student evaluation of teaching instrument really measure instructors' teaching effectiveness? An econometric analysis of students' perceptions in economics courses. *Econ. Anal. Policy* **2014**, *44*, 156–168. [CrossRef]
13. Darwin, S. Moving beyond face value: Re-envisioning higher education evaluation as a generator of professional knowledge. *Assess. Eval. High. Educ.* **2012**, *37*, 733–745. [CrossRef]
14. Clayson, D.E. Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature. *J. Mark. Educ.* **2008**, *31*, 16–30. [CrossRef]
15. Galbraith, C.S.; Merrill, G.B.; Kline, D.M. Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Res. High. Educ.* **2012**, *53*, 353–374. [CrossRef]
16. Friborg, O.; Rosenvinge, J.H. A comparison of open-ended and closed questions in the prediction of mental health. *Qual. Quant.* **2013**, *47*, 1397–1411. [CrossRef]

17. Ulmeanu, M.-E.; Doicin, C.-V.; Spânu, P. Comparative Evaluation of Sustainable Framework in STEM Intensive Programs for Secondary and Tertiary Education. *Sustainability* **2021**, *13*, 978. [CrossRef]
18. Arnon, S.; Reichel, N. Closed and Open-Ended Question Tools in a Telephone Survey About "The Good Teacher": An Example of a Mixed Method Study. *J. Mix. Methods Res.* **2009**, *3*, 172–196. [CrossRef]
19. Mendes, P.M.; Thomas, C.R.; Cleaver, E. The meaning of prompt feedback and other student perceptions of feedback: Should National Student Survey scores be taken at face value? *Eng. Educ.* **2011**, *6*, 31–39. [CrossRef]
20. Richardson, J.T.E.; Slater, J.B.; Wilson, J. The National Student Survey: Development, findings and implications. *Stud. High. Educ.* **2007**, *32*, 557–580. [CrossRef]
21. Deeley, S.J.; Fischbacher-Smith, M.; Karadzhov, D.; Koristashevskaya, E. Exploring the 'wicked' problem of student dissatisfaction with assessment and feedback in higher education. *High. Educ. Pedagog.* **2019**, *4*, 385–405. [CrossRef]
22. Mackay, J.R.D.; Hughes, K.; Marzetti, H.; Lent, N.; Rhind, S.M. Using National Student Survey (NSS) qualitative data and social identity theory to explore students' experiences of assessment and feedback. *High. Educ. Pedagog.* **2019**, *4*, 315–330. [CrossRef]
23. Langan, A.M.; Scott, N.; Partington, S.; Oczujda, A. Coherence between text comments and the quantitative ratings in the UK's National Student Survey. *J. Furth. High. Educ.* **2017**, *41*, 16–29. [CrossRef]
24. Hammonds, F.; Mariano, G.J.; Ammons, G.; Chambers, S. Student evaluations of teaching: Improving teaching quality in higher education. *Perspect. Policy Pract. High. Educ.* **2017**, *21*, 26–33. [CrossRef]
25. Cheng, J.H.; Marsh, H.W. National Student Survey: Are differences between universities and courses reliable and meaningful? *Oxf. Rev. Educ.* **2010**, *36*, 693–712. [CrossRef]
26. Douglas, J.; Douglas, A.; Barnes, B. Measuring student satisfaction at a UK university. *Qual. Assur. Educ.* **2006**, *14*, 251–267. [CrossRef]
27. Arthur, L. From performativity to professionalism: Lecturers' responses to student feedback. *Teach. High. Educ.* **2009**, *14*, 441–454. [CrossRef]
28. McClain, L.; Gulbis, A.; Hays, D. Honesty on student evaluations of teaching: Effectiveness, purpose, and timing matter! *Assess. Eval. High. Educ.* **2017**, *2938*, 1–17. [CrossRef]
29. Spooren, P.; Brockx, B.; Mortelmans, D. On the Validity of Student Evaluation of Teaching: The State of the Art. *Rev. Educ. Res.* **2013**, *83*, 598–642. [CrossRef]
30. Oberrauch, A.; Mayr, H.; Nikitin, I.; Bügler, T.; Kosler, T.; Vollmer, C. I Wanted a Profession That Makes a Difference—An Online Survey of First-Year Students' Study Choice Motives and Sustainability-Related Attributes. *Sustainability* **2021**, *13*, 8273. [CrossRef]
31. Mortelmans, D.; Spooren, P. A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educ. Stud.* **2009**, *35*, 547–552. [CrossRef]
32. Toland, M.D.; De Ayala, R.J. A Multilevel Factor Analysis of Students' Evaluations of Teaching. *Educ. Psychol. Meas.* **2005**, *65*, 272–296. [CrossRef]
33. MacFadyen, L.P.; Dawson, S.; Prest, S.; Gasevic, D. Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assess. Eval. High. Educ.* **2016**, *41*, 821–839. [CrossRef]
34. La Rocca, M.; Parrella, M.L.; Primerano, I.; Sulis, I.; Vitale, M.P. An integrated strategy for the analysis of student evaluation of teaching: From descriptive measures to explanatory models. *Qual. Quant.* **2017**, *51*, 675–691. [CrossRef]
35. Aphinyanaphongs, Y.; Fu, L.D.; Li, Z.; Peskin, E.R.; Efstathiadis, E.; Aliferis, C.F.; Statnikov, A. A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *J. Assoc. Inf. Sci. Technol.* **2014**, *65*, 1964–1987. [CrossRef]
36. Hofmann, T. Probabilistic Latent Semantic Indexing. *ACM SIGIR Forum* **2017**, *51*, 211–218. [CrossRef]
37. Wallace, B.C.; Trikalinos, T.A.; Lau, J.; Brodley, C.; Schmid, C.H. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinform.* **2010**, *11*, 55. [CrossRef] [PubMed]
38. Soto, A.; Kiros, R.; Kešelj, V.; Milios, E. Exploratory Visual Analysis and Interactive Pattern Extraction from Semi-Structured Data. *ACM Trans. Interact. Intell. Syst.* **2015**, *5*, 1–36. [CrossRef]
39. Kjellström, S.; Golino, H. Mining concepts of health responsibility using text mining and exploratory graph analysis. *Scand. J. Occup. Ther.* **2018**, *26*, 1–16. [CrossRef] [PubMed]
40. Sameh, A.; Kadray, A. Semantic Web Search Results Clustering Using Lingo and WordNet. *Int. J. Res. Rev. Comput. Sci.* **2020**, *1*, 71.
41. Morandi, A.; Limousin, M.; Sayers, J.; Golwala, S.R.; Czakon, N.G.; Pierpaoli, E.; Jullo, E.; Richard, J.; Ameglio, S. X-ray, lensing and Sunyaev-Zel'dovich triaxial analysis of Abell 1835 out to R_{200}. *Brain Behav. Immunit* **2011**, *25*, 1136–1142. [CrossRef]
42. Kleij, F.T.; Musters, P.A. Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Qual. Prefer.* **2003**, *14*, 43–52. [CrossRef]
43. Deneulin, P.; Bavaud, F. Analyses of open-ended questions by renormalized associativities and textual networks: A study of perception of minerality in wine. *Food Qual. Prefer.* **2016**, *47*, 34–44. [CrossRef]
44. Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. Structural Topic Models for Open-Ended Survey Responses. *Am. J. Political Sci.* **2014**, *58*, 1064–1082. [CrossRef]
45. Krosnick, J.A. Questionnaire design. In *The Palgrave Handbook of Survey Research*; Springer International Publishing: Cham, Switzerland, 2018; pp. 439–455. [CrossRef]

46. Reja, U.; Manfreda, K.L.; Hlebec, V.; Vehovar, V. Open-ended vs. close-ended questions in web questionnaires. *Dev. Appl. Stat.* **2003**, *19*, 159–177.
47. Mossholder, K.W.; Settoon, R.P.; Harris, S.G.; Armenakis, A.A. Measuring Emotion in Open-ended Survey Responses: An Application of Textual Data Analysis. *J. Manag.* **1995**, *21*, 335–355. [CrossRef]
48. Ang, C.S.; Lee, K.F.; Dipolog-Ubanan, G.F. Determinants of First-Year Student Identity and satisfaction in higher education: A quantitative case study. *SAGE Open* **2019**, *9*, 1–13. [CrossRef]
49. Awang, M.M.; Kutty, F.M.; Ahmad, A.R. Perceived Social Support and Well Being: First-Year Student Experience in University. *Int. Educ. Stud.* **2014**, *7*, 261–270. [CrossRef]
50. Grant-Vallone, E.; Reid, K.; Umali, C.; Pohlert, E. An Analysis of the Effects of Self-Esteem, Social Support, and Participation in Student Support Services on Students' Adjustment and Commitment to College. *J. Coll. Stud. Retent. Res. Theory Pract.* **2003**, *5*, 255–274. [CrossRef]
51. Wilcox, P.; Winn, S.; Fyvie-Gauld, M. It was nothing to do with the university, it was just the people: The role of social support in the first-year experience of higher education. *Stud. High. Educ.* **2005**, *30*, 707–722. [CrossRef]
52. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
53. Wang, S.; Manning, C.D. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Korea, 8 July 2012; pp. 90–94.
54. Onan, A.; Korukoğlu, S.; Bulut, H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Syst. Appl.* **2016**, *57*, 232–247. [CrossRef]
55. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
56. Lau, J.H.; Newman, D.; Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 530–539.
57. Watson, S. Closing the feedback loop: Ensuring effective action from student feedback. *Tertiary Educ. Manag.* **2010**, *9*, 145–157. [CrossRef]
58. Hettrick, S. What's wrong with computer scientists? *Softw. Sustain. Inst.* **2013**. Available online: https://www.software.ac.uk/blog/2016-10-06-whats-wrong-computer-scientists (accessed on 21 December 2021).
59. Why Your Computer Science Degree Won't Get You a Job. Available online: https://targetjobs.co.uk/career-sectors/it-and-technology/advice/323039-why-your-computer-science-degree-wont-get-you-an-it-job (accessed on 22 December 2021).
60. Ramsden, P. A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Stud. High. Educ.* **1991**, *16*, 129–150. [CrossRef]
61. Silveira, N.; Dozat, T.; De Marneffe, M.C.; Bowman, S.R.; Connor, M.; Bauer, J.; Manning, C.D. A gold standard dependency corpus for English. In Proceedings of the LERC, Reykjavik, Iceland, 26–31 May 2014.
62. Wissler, L.; Almashraee, M.; Monett, D.; Paschke, A. The gold standard in corpus annotation. In Proceedings of the IEEE GSC, Passau, Germany, 26–27 June 2014. [CrossRef]